DOCUMENT RESUME

ED 091 435 TM 003 649

AUTHOR Haladyna, Thomas M.
TITLE An Investigation of Full-And Subscale Reliabilities of Criterion-Referenced Tests.
PUB DATE [Apr 74]
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; Sampling; Standard Error of Measurement; Test Construction; *Test Reliability
IDENTIFIERS Classical Test Theory; Subscale Reliability; Variance (Statistical)

ABSTRACT

Classical test theory has been rejected for application to criterion-referenced (CR) tests by most psychometricians due to an expected lack of variance in scores and other difficulties. The present study was conceived to resolve the variance problem and explore the possibility that classical test theory is both appropriate and desirable for some types of CR tests. Both a rationale and empirical evidence were offered to support the practice of suing unrestricted samples to estimate full- and subscale reliabilities of CR tests using classical procedures. However, reservations were expressed concerning the reliability of these subscales. (Author)

AN INVESTIGATION OF FULL- AND SUBSCALE RELIABILITIES OF CRITERION-REFERENCED TESTS

Thomas M. Haladyna

Southern Illinois University at Carbondale

A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

AN INVESTIGATION OF FULL- AND SUBSCALE RELIABILITIES OF CRITERION-REFERENCED TESTS

Thomas M. Haladyna

Southern Illinois University at Carbondale

Abstract

Classical test theory has been rejected for application to criterion-referenced (CR) tests by most psychometricians due to an expected lack of variance in scores and other difficulties. The present study was conceived to resolve the variance problem and explore the possibility that classical test theory is both appropriate and desirable for some types of CR tests. Both a rationale and empirical evidence were offered to support the practice of using unrestricted samples to estimate full- and subscale reliabilities of CR tests using classical procedures. However, reservations were expressed concerning the reliability of these subscales.

AN INVESTIGATION OF FULL- AND SUBSCALE RELIABILITIES OF CRITERION-REFERENCED TESTS

Thomas M. Haladyna

Southern Illinois University at Carbondale

Considerable efforts have been directed toward the conceptualization and estimation of reliability of tests designed specifically for criterion-referenced (CR) measurement (e.g. Livingston, 1971; Ozenne, 1972; Hambleton and Novick, 1973). All of these efforts have stemmed from a rejection of classical test theory, and the grounds for this rejection have been either logical or statistical. However, this rejection may be premature for several reasons. The purpose of this study was to examine the concept of reliability as it applies to CR measurement and to determine if and how classical test theory may be used to study the reliability of CR tests.

The Nature of CR Measurement

Norm-referenced (NR) measurement involves the making of comparisons among examinee's test scores for the purpose of determining honors or grades and selecting, placing, or grouping persons. In these types of situations, classical test theory, as described by Lord and Novick (1968) and Nunnally (1967) has typically been employed. CR measurement involves the comparison of examinee's scores with an absolute standard for the purpose of determining whether or not the students have achieved at or beyond a desirable level.

Beyond this basic distinction between CR and NR measurement, a wide variety of definitions have been employed to characterize CR measurement (see Hambleton and Novick, 1973). Most have agreed that any CR test is instructional objective-based and that a criterion level is determined for the purpose of deciding which examinees have or have not mastered an objective or set of objectives. However, an examination of instructional objectives,

as they presently exist, reveals at least four different types, and each type requires a different kind of CR test.

Type 1. Often students are asked to perform or produce a result which can be directly observed and thereby verified. For example: Given one hour and materials as demonstrated in class, the student will construct a batik drawing on 8 1/2 by 11 inch poster paper which employs two different media and three primary colors. The product to be evaluated can easily be judged as completed within the specified time limit; consequently, the reliability of this type of performance or product assessment can be defined as the degree of concurrence among judges. Since the performances or products are directly viewed, this concurrence is perceived to be consistently high.

Type 2. At times inferences must be made regarding performances or products. Judges may be called upon to rate or rank these performances or products with respect to form, style, technical excellence, creativity, or a host of similar characteristics. This type of instructional objective may [illegible] to plays, essays, poetry, sculture, gymnastic competition, similar curricular and extracurricular activities. In Type 2 objectives, inferences may lead to a greater lack of agreement among judges. Reliability is estimated using familiar intraclass correlation techniques.

Type 3. In some instances, the attainment of a single objective is of great importance. In this situation, a single test may be constructed which measures the degree of attainment of that objective. For example: Given statements representing examples of physical changes, the student will identify with 80% accuracy which of six possible changes has occurred. Test items from such a test can be dichotomously scored, and each item is in itself a measure of that objective. The reliability of this apparently homogeneous test appears to be appropriately estimated using internal consistency techniques such as KR-20. One problem with this approach to reliability, as pointed out by Popham and Husek (1969), is the possibility

that the variability of postinstruction test scores is so drastically reduced that traditional reliability estimates, which rely on variance, are useless.

Type 4. Millman (1973) has described a type of CR test which measures performance with respect to a universe of interrelated test items. The phrase "domain-reference" has been used to denote this type of CR test, and the procedures for constructing a domain-referenced test are: (a) an achievement domain is hypothesized in terms of related instructional objectives; (b) the objectives are organized into subsets representing various regions of the domain; (c) test items are written to measure acquisition of each objective; and (d) tests are constructed by sampling items from these subsets. These procedures are similar to those described by Cronbach and Meehl (1955) for validating psychological constructs and to those described by Nunnally (1967) in his treatment of classical test theory. Further, these procedures are also consonant with practices employed in developing tests with high factorial or sampling validity, very desirable forms of content validity.

The Use of CR Tests in Instruction. As a result of mastery type learning paradigms suggested by Carroll (1963) and Bloom (1968), there has been an increasing emphasis on individual instruction which features a careful specification of intended instructional outcomes in the form of instructional objectives, preassessment, instruction, and postassessment. If a satisfactory level of performance is reached, the student continues in a sequence of instruction. When a student fails to surpass the criterion level following instruction, he may be redirected to study his instructional objectives and related materials or seek other remedial help of a nonspecific nature. An alternative instructional strategy might be to diagnose learning difficulties in terms of regions of the domain. The information obtained

from subscales of domain-referenced tests may be used to offer specific remedial instruction. It is with respect to the decision-making both at full and subscales that reliability becomes important.

CR Reliability

The problem which confronts the instructor who uses CR tests in mastery instruction is whether or not a student has reached or surpassed the criterion level, particularly those persons whose scores fall close to or at the criterion level. For these persons, two types of errors may occur: (a) true nonmastery students may be classified as mastery students and (b) true mastery students may be classified as nonmastery. One way of combatting these errors is to set a confidence interval around the criterion level. Those falling in a critical region above the criterion level may be classified as mastery, while those falling in the critical region below the criterion level may be classified as nonmastery. Those falling in the confidence interval have questionable status due to the closeness of their scores to the criterion level. Consequently, subscale diagnosis may yield information about regions which have not been mastered. Thus specific remedial instruction may occur, a retest may be administered over that specific region, as measured by a subscale, and improvement in performance may result in the classification of the student in the mastery group.

Popham and Husek (1969), among others, have rejected classical test theory for analyzing CR tests primarily due to the suspicion that the variance of postinstruction scores is too restricted. Since classical test theory is largely dependent on large test score variance, item discrimination indexes and reliability estimates would be attenuated. It has also been argued that since the purposes of CR testing are quite different from those of NR testing, a set of procedures quite independent of classical theory is necessary. Only recently have these arguments been challenged (Klein and

Kosecoff, 1973; Woodson, 1974; Haladyna, 1974). First restricted variation of CR test scores of the domain-referenced type may occur only when instruction is highly effective. Second, this restriction in range of CR test scores following instruction may be due to the selection of examinees rather than some intrinsic and unique characteristic of CR tests. Finally, the primary difference between CR and NR tests may not be in their construction and analysis but rather in how the tests are used. That is, in the final analysis, the estimation of true scores is important in both CR and NR tests. Thus the central goal is the determination of the degree of error, and this is accomplished through the use of the traditional reliability coefficient.

In classical theory, the magnitude of item discrimination indexes is functionally related to reliability (Scott, 1961; Guilford, 1965). New techniques have been proposed for CR measurement based on group differences, and several studies have indicated that a close correspondence exists between classical estimates of CR item discrimination and these new group difference techniques (Helmstadter, 1972; Haladyna, 1974). If classical indexes of item discrimination adequately measure discrimination of CR test items, can classical reliability estimates be used in CR tests?

Another issue that has been raised in connection with the applicability of classical test theory for CR tests is the internal consistency of CR tests which represent numerous regions, that is, multiscaled tests may not be highly internally consistent (Shavelson, Block, and Ravitch, 1972). Therefore, internal consistency reliability is said to be inappropriate for these multiscaled tests. However, if all regions have a commonality with the domain; the subscales representing these regions may be highly internally consistent as well as highly intercorrelated, and the fullscale

homogeneity may also be reasonably high. The degree of reliability obtained through the use of unrestricted samples (mastery and nonmastery examinees) for both full- and subscales is an empirical question. Consequently, the following questions were formulated:

1. When unrestricted samples containing both mastery and nonmastery examinees are employed, can internal consistency techniques provide adequate and useful estimates of reliability?

2. What effects do unrestricted samples have on the estimates of errors of measurement?

3. To what degree does conceptually organizing instructional units into regions lead to internally consistent subscales? To what degree are these regions, as represented by subscales, interrelated?

4. Do the number of regions and associated subscales attenuate the fullscale internal consistency of these domain-referenced tests?

5. Does the length of subscales predictably affect the magnitudes of internal consistency measures when unrestricted samples are used?

METHOD

Subjects. Nearly 180 students enrolled in an undergraduate level measurement and evaluation course were administered CR achievement tests as part of normal instruction. These students were mostly females; juniors and seniors; and special education majors. Their grade point averages and American College Test scores were similar to those of the university population.

Construction of Achievement Tests. Instructional objectives were classified into seven basic units; student achievement in three of these units was evaluated through the use of CR tests of the domain-referenced type. Unit One consisted of concepts related to the construction and use

of teacher-made tests; Unit Two was related to basic statistical concepts; Unit Three was focused on standardized tests. Unit One had six regions; Unit Two had four regions; and Unit Three had five regions. Items were constructed or selected from existing item files to relate to instructional objectives which represented various regions. Items were randomly assigned to one of three parallel test forms; these forms varied in length from 40 to 50 items. Subscales varied in length from two to 17 items depending upon the number of objectives in a particular region.

Procedure. Mastery learning was explained to all students both orally and in writing. Every student was pretested using one form and tested following instruction using another form. A third form was used for retesting when mastery was not demonstrated immediately following instruction. The criterion level was set at 70%, and in rare instances when students failed retests, diagnosis was done by regions and students were given remedial instruction and retests based on subscale information.

RESULTS AND DISCUSSION

Fullscale and subscale homogeneities were estimated using the KR-20 formula, and intercorrelations were computed among subscales. Means, standard deviations, standard errors of measurement, and homogeneity estimates for all units and forms for both restricted and unrestricted samples are presented in Table 1. Sample sizes were not proportionate for pre- and post-instruction samples due to the fact that the latter sample included retests for students failing the postinstruction test. Differences between pre- and postinstruction test scores, regardless of forms, indicated highly effective instruction. Not only were tests of differences statistically significant ($p < .001$), but the magnitudes of these differences were considerable.

1. Traditional NR test practices include item and test analyses following instruction. Since time to learn is held constant and learning rates vary, a large variance is observed in postinstruction test scores. In mastery instruction, where time to learn is allowed to vary, the variability of test scores is believed to be low, and this restriction would lead to invalid estimates of reliability. Homogeneity estimates reported in Table 1 confirm this suspicion; postinstruction homogeneity estimates varied from .31 to .72. When unrestricted samples were used, a predictable increase in these estimates resulted. KR-20 coefficients ranged from .69 to .89 with a median of .84. The magnitudes of these increases in homogeneity estimates, which resulted from using unrestricted samples, ranged from .1 to .41 with a median increase of .25. There was a direct correspondence between increases in variance and increases in reliability estimates.

Since a reliability coefficient is a descriptive index, it seems that using unrestricted samples consisting of both mastery and nonmastery examinees offers a better description of the degree of reliability possessed by these domain-referenced CR tests. The homogeneity estimates were satisfactorily high, and variance did not seem to be an issue.

2. Errors of measurement are said to be constant regardless of the variability of test scores for any particular sample. Consequently, one might expect standard errors of measurement to be constant across pre-, postinstruction, and combined samples. The results reported in Table 1 confirm this hypothesis. If standard errors of measurement are to be used for the setting of confidence intervals in order to permit useful and accurate decisionmaking, then these standard errors may be obtained from any sample. In any instructional setting where domain-referenced tests are employed, the standard error of measurement could be estimated from

Table 1

Means, Standard Deviations, Standard Effors of Measurement, and Homogeneity Estimates for Restricted and Unrestricted Samples Across All Forms and Units

	Preinstruction				Postinstruction				Combined			
Unit One	Mean	S.D.	SEM	KR-20	Mean	S.D.	SEM	KR-20	Mean	S.D.	SEM	KR-20
Preinstruction	18.8	5.6	2.9	.73	14.9	7.9	2.7	.88	25.6	5.3	3.0	.67
Postinstruction	34.0	4.7	2.5	.72	29.6	3.8	2.6	.52	31.8	4.0	2.8	.52
Combined Groups	25.8	8.7	2.8	.89	26.8	5.9	2.8	.77	28.8	5.6	2.9	.72
Unit Two												
Preinstruction	17.4	4.5	2.8	.60	14.5	3.3	2.7	.33	13.6	5.7	2.7	.77
Postinstruction	28.7	4.0	2.6	.58	27.4	3.3	2.5	.43	24.2	4.7	2.7	.68
Combined Groups	23.4	7.0	2.8	.84	24.2	6.5	2.6	.84	20.9	7.2	2.8	.86
Unit Three												
Preinstruction	21.2	5.8	2.8	.77	20.1	4.6	2.8	.62	20.8	4.3	2.6	.64
Postinstruction	28.8	4.0	2.6	.59	30.0	4.1	2.6	.61	26.0	2.6	2.2	.31
Combined Groups	26.0	6.0	2.7	.80	27.1	7.5	2.6	.88	24.1	4.2	2.3	.69

preinstruction test results, and the standard error could then be applied to postinstruction test scores to decide who has clearly passed, who has clearly failed, and who is in need of specific remediation.

3. The homogeneity estimates for full- and subscales as well as intercorrelations among subscales are presented in Tables 2, 3, and 4 for the three units of instruction respectively. In Unit One, there were fewer items per subscale due to the large number (six) of subscales. Homogeneity estimates ranged from .11 to .76 with a median of .45. Despite these low to moderate homogeneities, intercorrelations were often as high as reliabilities of these subscales permitted. In the few instances where correlations were low, scales typically involved consisted of fewer than five items. Most of these correlations among subscales were statistically significant, but more importantly, the magnitudes were consistently high. Since reliability attenuates correlation, when these correlations were corrected for attenuation, coefficients often approached or exceeded one. The latter instances point to situations where reliability may have been underestimated. Thus the six subscales of Unit One appear to have much in common despite the obvious uniqueness of each subscale and corresponding region. In Unit Two, intercorrelations both before and after correction for attenuation were extremely high. Intercorrelations of subscales in Unit three were also high with exceptions in the third form. In form C, the low variance of test scores for the unrestricted sample appeared to yield corresponding low full- and subscale homogeneity estimates as well as low intercorrelations among subscales. It appears that a conceptual organization of instructional objectives and related test items leads to relatively homogeneous subscales, and that these subscales are highly related. Further, these correlations among subscales appear to be limited only to the degree of the reliabilities of the subscales involved in each relationship.

Table 2

Intercorelations Among Subscales for Unit One, Forms A, B, and C

Form A	1	2	3	4	5	6	Number of Items
1. Introductory Concepts	(36)	126	111	133	117	106	8
2. Test Planning	60	(64)	106	102	107	92	10
3. Selected Response Tests	50	64	(57)	106	100	90	9
4. Constructed Response Tests	48	50	48	(36)	111	101	6
5. Posttest Activities	61	75	65	59	(76)	99	12
6. Grading Practices	47	56	51	46	65	(56)	3
Form B							
1. Introductory Concepts	(19)	120	130	105	99	103	8
2. Test Planning	41	(62)	107	81	92	84	7
3. Selected Response Tests	41	61	(53)	78	115	142	4
4. Constructed Response Tests	32	44	39	(49)	89	14	10
5. Posttest Activities	24	40	46	34	(31)	130	13
6. Grading Practices	15	22	34	03	24	(11)	3
Form C							
1. Introductory Concepts	(23)	117	74	86	109	109	7
2. Test Planning	30	(30)	65	129	86	11	8
3. Selected Response Tests	25	25	(48)	98	65	-35	15
4. Constructed Response Tests	27	45	44	(42)	81	02	5
5. Posttest Activities	37	33	32	37	(49)	41	11
6. Grading Practices	22	02	-10	01	12	(18)	3

[1] KR-20 estimates appear in parentheses, correlations appear below the diagonalof reliability estimates, correlations corrected for attenuation appear above the diagonal, all decimals have been omitted.

Table 3

Intercorrelations Among Subscales for Unit Two, Forms A, B, and C[1]

Form A	1	2	3	4	Number of items
1. Scales of Measurement	(67)	86	91	105	6
2. Statistical Concepts	57	(66)	82	98	9
3. Correlations and Prediction	65	58	(76)	120	11
4. Validity and Reliability	42	39	51	(23)	15
Form B					
1. Scales of Measurement	(62)	91	74	127	5
2. Statistical Concepts	62	(76)	89	122	9
3. Correlation and Prediction	50	67	(74)	119	11
4. Validity and Reliability	38	41	39	(15)	15
Form C					
1. Scales of Measurement	(60)	96	93	88	6
2. Statistical Concepts	62	(70)	80	92	9
3. Correlation and Prediction	53	49	(54)	84	9
4. Validity and Reliability	55	62	50	(64)	17

[1] KR-20 estimates appear in parentheses, correlations appear below the diagonal of reliability estimates, correlations corrected for attenuation appear above the diagonal of reliability estimates, all decimals have been omitted.

Table 4

Intercorrelations Among Subscales for Unit Three, Forms A, B, and C[1]

Form A	1	2	3	4	5	Number of items
1. Historical Background	(51)	37	39	59	50	6
2. Cognitive Tests	22	(66)	84	69	91	8
3. Affective Tests	17	42	(38)	95	94	2
4. Testing Programs	28	37	39	(44)	82	10
5. Interpreting and Reporting	27	57	44	42	(58)	17
Form B						
1. Historical Background	(52)	54	117	45	52	8
2. Cognitive Tests	30	(61)	137	106	101	10
3. Affective Tests	30	38	(13)	155	115	4
4. Testing Programs	26	68	45	(67)	101	8
5. Interpreting and Reporting	34	70	36	73	(79)	15
Form C						
1. Historical Background	(17)	13	-15	18	47	3
2. Cognitive Tests	04	(53)	-31	92	93	12
3. Affective Tests	-02	-09	(15)	-103	-61	3
4. Testing Programs	04	37	-22	(30)	96	5
5. Interpreting and Reporting	14	50	-18	40	(56)	17

[1] KR-20 estimates appear in parentheses, correlations appear below the diagonal of reliability estimates, correlations corrected for attenuation appear above the diagonal, all decimals have been omitted.

4. No relationship was observed between the number of subscales for any test form and the fullscale homogeneity estimates. The suspicion that any domain-referenced test which contains a great many subscales may have low internal consistency was not confirmed by the results of this study. A more serious byproduct of having too many subscales is the limitation of the number of items for each subscale.

5. From classical test theory, a high relationship is normally expected between the number of items in any scale and the homogeneity estimate for that scale. This relationship was not observed in the results of this study. Instead, the correlation between homogeneity estimates and number of items in scales was slightly positive and non-significant. Scales which possessed low homogeneity estimates also had restricted variances. While low reliability might be a plausible assumption about these scales, low reliability estimates also result when instruction has been ineffective or the items lack content validity (items did not measure what was taught). Despite the unexpected lack of relationship between subscale length and reliability, these estimates were seldom high enough to suggest a high degree of confidence. More importantly, the setting of confidence intervals about a criterion level for these subscales for the purpose of decisionmaking appears to be an extremely risky venture when considering the large standard errors of measurement which exist for these subscales. The limiting factor ultimately is the number of test items employed. As Hambleton and Novick (1973) have observed, the particular problem of deciding upon the number of items for any subscale has not yet been satisfactorily resolved. If decisionmaking is to be done at the fullscale level, the standard error of measurement, which can be estimated from any sample, can be usefully employed. When decisionmaking is done at the subscale level, it seems desirable to employ fewer subscales and

maximize the number of test items for each subscale.

One alternative to the use of homogeneity estimates at the fullscale level exists. Reliability can be estimated for various subscales through the use of technique where the subscales are treated as a linear combination (Nunnally, 1967). Reliabilities were estimated for all units and forms using both the KR-20 formula and the linear combination formula. As shown in Table 5, nearly identical reliability coefficients resulted. Thus it seems that these KR-20 coefficients are reasonably accurate estimates of reliability despite the obvious multidimensional composition of each test, and the belief that multiscaled tests would lack high internal consistency was not supported by these data.

The present study has been concerned with the usability of classical test theory for CR tests. In the context of systematic, mastery-based instruction, a logical rationale and empirical evidence has been offered to support the use of classical theory for estimating reliability through the use of internal co. .stency formuale. The problem that persists is the reliability of subscales, and more specifically, how can decisionmaking be improved at the subscale level. Since a number of new approaches to CR test reliability of the domain-referenced type have been proposed, it would be interesting to investigate subscale reliability using some of these new approaches.

Table 5

Comparison of Reliability Estimates Computed Two Different Ways for All Units and Forms

	Unit One		Unit Two		Unit Three	
	KR-20	Linear Comb.	KR-20	Linear Comb.	KR-20	Linear Comb.
Form A	.895	.961	.842	.845	.801	.817
Form B	.770	.771	.835	.867	.875	.880
Form C	.725	.740	.863	.871	.686	.699

REFERENCES

Bloom, B. S. Learning for Mastery. Evaluation Comment, 1968, 1, No. 1.

Carroll, J. B. A model for school learning. Teachers College Record, 1963, 64, 723-733.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Guilford, J. P. Fundamental statistics in education and psychology. New York: McGraw-Hill, 1965.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, in press.

Hambleton, R. K. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Helmstadter, G. C. A comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance oriented instruction. Paper presented at the annual meeting of the American Psychological Association, Honolulu, 1972.

Klein, S. P. & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. Eric TM Report 26, September, 1973.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lo[illegible] F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Ozenne, D. G. Toward an evaluative methodology for criterion-referenced measures. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Popham, W. J. & Husek, T. R., Implications of criterion-referenced measurement. Journal of Educational Measurement. 1969, 6, 1-9.

Scott, W. A. Measures of test homogeneity. Educational and Psychological Measurement, 1960, 20, 751-760

Shavelson, R., Block, J. & Ravitch, M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 9, 133-138.

Woodson, C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.